

PR #2379 完整报告

radixark/miles

docs: point the TB2 agent-server eval at the public harbor branch

合并时间: 2026-08-12 02:14

原文链接: <http://prhub.com.cn/radixark/miles/pull/2379>

执行摘要

- 一句话: TB2 eval 示例改指公共 harbor 分支
- 推荐动作: 值得快速浏览: 它是一个典型的『文档可访问性』修复, 展示了在公共示例中移除私有仓库依赖的规范化做法, 并与既有的 swe-agent-harbor 示例形成统一体系。若你的团队需要维护对外公开的实验示例, 可借鉴作者『切换指针前逐项核对目标分支能力』的自证流程。

功能与动机

示例放在公共仓库却要求读者从私有仓库 radixark/harbor-private 的 shi/rebase-on-upstream-v0.7.0 分支拉取 agent server, 组织外用户无法克隆, 文档与示例的可复现性矛盾直接阻碍外部用户使用。与此同时, Miles 的 agent server 集成已存在于公共仓库 harbor-framework/harbor 的 harbor-miles-v0.20.0 分支, 且 examples/swe-agent-harbor-docker 与 swe-agent-harbor-daytona 已经指向该分支, 因此本 PR 将示例指针统一到同一公共分支, 既修复可访问性问题, 也保持示例之间的一致性。作者在 PR 描述中逐项核对了目标分支上示例依赖的 miles_agent_server.py 参数、poll-steps 包装器、/api/sessions、89 任务的 registry.json 以及 harbor datasets download --export, 确保命令可直接生效。

实现拆解

1. 更新 eval_tb_deepseek_v4_pro.py 的头部文档与参数帮助: 模块 docstring 中『harbor-private 的 shi/rebase-on-upstream-v0.7.0 分支』改为『harbor 的 harbor-miles-v0.20.0 分支』; --registry-path 的用法示例由 /path/to/harbor-private/registry.json 改为 /path/to/harbor/registry.json; argparse 中该参数的 help 文本同步改为 Path to harbor's registry.json。
2. 更新 README.md 四处内容: Prerequisites 第 1 步的分支链接改为 harbor-framework/harbor 的 harbor-miles-v0.20.0 分支; 第 4 步 registry.json 来源由 harbor-private checkout 改为 that harbor checkout; Run 章节的 --registry-path 命令参数示例同步替换; Tuning 章节中 --max-seq-len 65536 的描述由 harbor-private branch's poll-steps wrapper 改为 harbor branch's poll-steps wrapper, 并更新文末链接脚注。
3. 验证与配套说明: 作者在目标分支上核对了示例依赖的全部能力, 但本 PR 仅变更文档与 docstring, 无测试文件增减; README 中的 DeepSeek-V4-Pro 评测数据仍来自旧私有分

支，未在新分支上复测，这一点已在 PR 描述中如实声明。

关键文件：

- `examples/experimental/eval/terminal_bench_via_agent_server/eval_tb_deepseek_v4_pro.py`（模块 评测示例；类别 `source`；类型 `documentation`；符号 `parse_args`）：示例的核心客户端脚本；模块 `docstring` 与 `--registry-path` 参数帮助文本从私有 `harbor-private` 分支改为公共 `harbor-framework/harbor` 的 `harbor-miles-v0.20.0` 分支，虽然无逻辑变更，但这是外部读者最先看到且必须准确的入口。
- `examples/experimental/eval/terminal_bench_via_agent_server/README.md`（模块 评测示例；类别 `docs`；类型 `documentation`）：面向用户的主文档，集中承载分支链接、环境准备步骤、运行命令与调优说明；本次把 `harbor-private checkout` 全部替换为公共 `harbor` 仓库，并更新链接脚注，是外部用户能否复现的关键依赖信息。

关键符号： `parse_args`

关键源码片段

`examples/experimental/eval/terminal_bench_via_agent_server/eval_tb_deepseek_v4_pro.py`

示例的核心客户端脚本；模块 `docstring` 与 `--registry-path` 参数帮助文本从私有 `harbor-private` 分支改为公共 `harbor-framework/harbor` 的 `harbor-miles-v0.20.0` 分支，虽然无逻辑变更，但这是外部读者最先看到且必须准确的入口。

```
"""Evaluate DeepSeek-V4-Pro on Terminal-Bench 2.0 via a running ``miles_agent_server``.
```

纯客户端脚本：不涉及训练、Ray 或 Megatron，只负责向已运行的 agent server 并发 `POST ``/run``` 请求，并聚合每次 `trial` 的结果。

```
# 本次指针变更：原文档要求使用私有仓库 radixark/harbor-private 的  
# shi/rebase-on-upstream-v0.7.0 分支，组织外用户无法克隆；现改为公共仓库  
# harbor-framework/harbor 的 ``harbor-miles-v0.20.0`` 分支，与既有  
# swe-agent-harbor-docker / swe-agent-harbor-daytona 示例保持一致。  
The agent server is expected to be the one shipped in harbor's  
``harbor-miles-v0.20.0`` branch (see that repo's README for setup), running  
with ``$HARBOR_TASKS_DIR`` populated with the 89 Terminal-Bench 2.0 tasks.  
The server is dataset-agnostic; this script just resolves the 89 task names  
from ``registry.json`` and dispatches one ``/run`` request per (task, trial) pair.
```

Usage::

```
export DEEPSEEK_API_KEY=<your-key>  
python eval_tb_deepseek_v4_pro.py \  
    --server-url http://localhost:8080 \  
    --registry-path /path/to/harbor/registry.json \  
    --n-trials-per-task 4 \  
    --max-concurrent 16
```

```
"""
```

```

import argparse
import asyncio
import json
import os
from collections.abc import Awaitable
from pathlib import Path
from time import monotonic
from typing import Any

import httpx
import polars as pl

def parse_args() -> argparse.Namespace:
    parser = argparse.ArgumentParser(
        description="Evaluate DeepSeek-V4-Pro on Terminal-Bench 2.0 via miles_agent_server.",
        formatter_class=argparse.ArgumentDefaultsHelpFormatter,
    )
    # --registry-path 的路径示例与帮助文本同步从 harbor-private 改为公共 harbor,
    # 用户可直接从公共 checkout 的 registry.json 读取 89 个 TB2 任务名。
    parser.add_argument(
        "--registry-path",
        type=Path,
        default=Path("registry.json"),
        help="Path to harbor's registry.json (used to read the 89 TB2 task names).",
    )
    # ... 其余参数 (--dataset-name、--n-trials-per-task、--max-concurrent 等) 保持原样
    return parser.parse_args()

```

评论区精华

本 PR 没有产生实质 review 讨论；唯一的审核意见是 Shi-Dong 的批准『LGTM, thanks!』，未提出任何问题或改动请求。作者在 PR 描述中自证的目标分支逐项核验（miles_agent_server.py 参数、poll-steps 包装器、/api/sessions、registry.json 的 89 个任务、harbor datasets download --export）为无 review 评论的顺利合入提供了依据。

- 暂无高价值评论线程

风险与影响

- 风险：
 1. 评测成绩与分支不一致：README 中的 DeepSeek-V4-Pro 数据是在旧私有分支上测得的，作者明确说明未在新分支复跑，外部用户在新分支上复现时结果可能与文档数字有差异。
 2. 外部依赖稳定性：文档链接依赖 harbor-framework/harbor 的 harbor-miles-v0.20.0 分支长期存在且内容不变，若上游调整分支名或代码，示例将失效。

3. 未端到端实跑：虽然作者逐项核对了文件与参数，但示例整体未在新分支上实际运行过，命令细节可能与真实运行存在偏差。
4. 影响面有限：不涉及训练、rollout、模型或核心库代码，回归风险主要停留在文档层。
 - 影响：影响范围限定在 `examples/experimental/eval/terminal_bench_via_agent_server` 目录，不涉及核心库。对用户：组织外部用户现在可以克隆公共 harbor 分支并按示例复现 TB2 评测，解决了『公开但不可复现』的矛盾。对团队：统一了多个示例对 harbor agent server 分支的指向，降低文档维护分散度。对系统：无运行时行为变化，纯文档与注释改动。
 - 风险标记：评测数据未在新分支复验，依赖外部仓库分支稳定性，文档示例未端到端实跑

关联脉络

- PR #2373 docs: NeMo-Gym server no longer needs a fork branch: 同属『文档可访问性』修复：移除对用户不可获取的分支或仓库依赖，改为指向可直接使用的公共来源。
- PR #2220 Add the GLM-5.2 744B x terminal-bench-2 Daytona example: 同一 Terminal-Bench 2.0 / harbor agent 生态的示例建设，共享对外公开的运行指引方向。
- PR #2274 refactor(openenv): move duplicated sandbox helpers into the TB2 recipe: TB2 相关基础设施复用，间接与本例的 harbor agent server 协作，同属 Terminal-Bench 评测工具链演进。